Combining Functional Data Registration and Factor Analysis

Cecilia Earls
Cornell University
and
Giles Hooker
Cornell University

June 8, 2015

Abstract

We extend the definition of functional data registration to encompass a larger class of registration models. In contrast to traditional registration models, we allow for registered functions that have more than one primary direction of variation. The proposed Bayesian hierarchical model simultaneously registers the observed functions and estimates the two primary factors that characterize variation in the registered functions. Each registered function is assumed to be predominantly composed of a linear combination of these two primary factors, and the function-specific weights for each observation are estimated within the registration model. We show how these estimated weights can easily be used to classify functions after registration using both simulated data and a juggling data set.

Keywords: Bayesian modeling, factor analysis, functional data, registration, variational Bayes

1 INTRODUCTION

This paper extends current functional data registration methods to encompass a broader family of registration models. Traditional registration methods are designed to eliminate all phase variability in a set of functions so that amplitude variability in the registered functions can be described by one primary functional direction. A simple example can be found in Figure 1. Here, each unregistered function, $X_i(t)$, in the center plot can be expressed as $X_i(t) = z_{1i} * f_1(t + c_i)$ where $f_1(t)$ is the primary direction of variation in the registered functions. After registration, these functions only

exhibit amplitude variability in the direction of $f_1(t)$ as seen in the illustration on the right. For a thorough discussion on the history of and current methods in functional registration, see Earls and Hooker [4].

We build on this traditional concept of functional registration, by considering unregistered functions for which eliminating phase variability in these functions results in registered functions that vary in two primary functional directions which we will denote $f_1(t)$ and $f_2(t)$. Allowing two primary functional directions of variation in the registered functions extends the use of functional registration to functional data sets such as that found in Figure 2. Here the composition of some of the registered functions includes a negative scaling of the second factor which confounds traditional approaches to registration. Considering these factors separately in the registration process is essential to eliminate phase variability in these functions.

The registration model presented here is an extension of our previous work in traditional functional registration in the framework of a Bayesian hierarchical model, Earls and Hooker [4]. In our previous work, we demonstrate this approach to functional registration not only allows for flexible modeling assumptions, but also results in estimates of registered functions that are similar to those determined by the best registration procedures currently available. Here we will extend this model to not only register functions with multiple directions of variation after registration, but also to perform factor analysis. For these models, approximate inference can be performed with an adapted variational Bayes algorithm that significantly reduces the computational time needed for initial estimates. Using these estimates are to initialize an MCMC sampling scheme eliminates the need for a burn-in period. Appendix B provides the details of this algorithm for the registration and factor analysis model. A complete discussion of the adapted variational Bayes (AVB) algorithm can be found in Earls and Hooker [4] where we also compare AVB estimates to those obtained by MCMC sampling for several data sets.

There is no previous work that combines registration and factor analysis; however, in Kneip and Ramsay [8], the authors also consider registration where the aligned functions are assumed to contain variation in more than one functional direction. In their paper, Kneip and Ramsay register functions using an iterative algorithm that updates the PCA decomposition used to register functions in each iteration. This model can be seen as an extension of the Procrustes method for traditional functional registration, Ramsay and Li [15] and Ramsay and Silverman [16]. Earls and

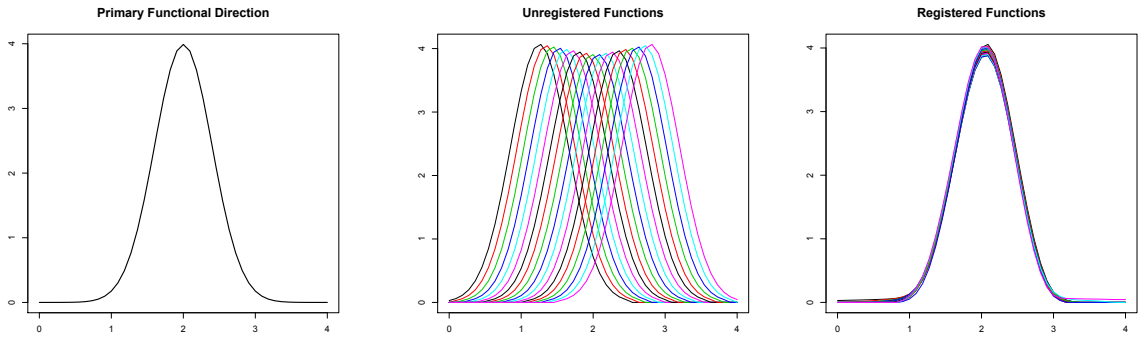

Figure 1: Example of Traditional Function Registration. **Left** The functional direction in which describes all variation in the registered functions, $f_1(t)$. **Center** Each unregistered function is a scaling of the primary functional direction that is shifted horizontally. These horizontal shifts account for the phase variability in the data. **Right** The functions after registration.

Hooker [4] demonstrates how our initial registration model improves upon the Procrustes method.

The basic organization of this paper is as follows. We present our model for registration and factor analysis in Section 2. In Section 3 we compare our model for functional alignment to one of the best traditional registration methods using two simulated data sets. In this section, we also show how functions can be grouped according to their estimated weights on each of the two factors. In Section 4, we apply this model to a juggling data set. Finally, a discussion can be found in Section 5.

2 FACTOR ANALYSIS MODEL FOR REGISTRATION AND GROUPING

2.1 Informative Precision Matrices for Functional Data Registration

We extend Earls and Hooker [4], on functional registration via Gaussian process models, to allow for more flexible assumptions in the structure of the registered functions. Using the classical definition of functional registration, in Earls and Hooker [4], we propose a registration model designed to register functions that once registered have little variation from one functional direction. While appropriate for many statistical analyses, this registration model does not adequately register

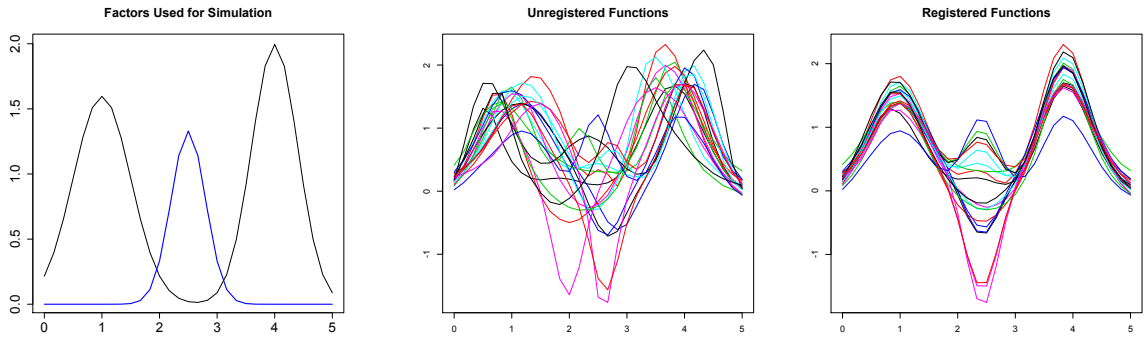

Figure 2: Example of Expanded Function Registration. **Left** The two functional directions that describe all variation in the registered functions, $f_1(t)$ and $f_2(t)$. **Center** Each unregistered function is composed of a linear combination of $f_1(t)$ and $f_2(t)$ with non-linear phase variation. **Right** The functions after registration.

functions in which there are more than one primary direction of variation in the registered functions. As we will show in Section 3, other registration methods based on this traditional definition of registration also tend to perform poorly when the registered functions are composed of more than one primary direction of variation.

Earls and Hooker [4] establishes that under the assumption that the registered functions vary insignificantly from one primary functional direction, the following data distribution is appropriate to register functions $X_i(t)$, $i = 1, \ldots, N$.

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t) \sim GP(z_{0i} + z_{1i} f_1(t), \gamma_1^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \tag{1}$$

where $X_i(h_i(t))$ is $X_i(t)$ registered under the warping function $h_i(t)$. The above covariance function, $\gamma_1^{-1}\Sigma(s,t)$, penalizes all variance from a scaling and vertical shifting of the primary functional direction, $f_1(t)$. In these models we will define γ_1 as a registration parameter that determines the severity of this penalty. This registration parameter is balanced by a penalty on the warping functions, $h_i(t), i = 1, \ldots, N$ that penalizes distance from the identity warping. For more information on this model see Earls and Hooker [4].

It is natural to extend this initial model to

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t), z_{2i}, f_2(t) \sim GP(z_{0i} + z_{1i} f_1(t) + z_{2i} f_2(t), \gamma_1^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \tag{2}$$

However, this distribution penalizes variation from the first and second functional directions

(factors), $f_1(t)$ and $f_2(t)$, equally. For most data, variation in one of the factors will exceed variation in the other factor. Accounting for this discrepancy in the statistical model for the registered functions not only provides a better registration, but also creates an identifiable relationship between the two factors. We will thus proceed with the following distribution for the registered functions.

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t), z_{2i}, f_2(t) \sim GP(z_{0i} + z_{1i}f_1(t) + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i} f_2(t), (\gamma_1+\gamma_2)^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T}$$

This distribution introduces separate penalties for 1) variation from the first functional direction, $\mathbf{f}_1$, controlled by registration parameter, γ_1, and 2) variation from a linear combination of $\mathbf{f}_1$ and $\mathbf{f}_2$ controlled by registration parameter γ_2, $\gamma_2 < \gamma_1$.

Before establishing the basis for the exact specification of the distribution above, we note here, as is common with functional data, that each unregistered function, $X_i(t)$, is assumed to be observed over a finite number of equally spaced time points, $\mathbf{t} = (t_1, \ldots, t_p)'$. Thus, given the above model, in practice we will proceed by using finite approximations to each function. In Earls and Hooker [3] we establish some theoretical properties of these types of approximations. The following finite-dimensional distribution is used in the final model in lieu of its infinite dimensional counterpart above. For $\mathbf{X}_i(\mathbf{h}_i) = (X_i(h_i(t_1)), \ldots, X_i(h_i(t_p)))'$, $i = 1, \ldots, N$,

$$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \quad \sim \quad N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \tag{3}$$

The underlying principle for both the registration and factor analysis model presented here and the basic registration model described in Earls and Hooker [4] is the use of *informative* priors in a Bayesian hierarchical model. The mean vectors and precision matrices used in the prior distributions of the registered functions, (1) and (3), for these models are selected to define the types of variation allowable for functions that are fully registered. Explicitly defining proper covariance relationships for registered functions in these prior distributions results in posterior estimates of the registered functions that are registered by warping the time domain of each unregistered function until the covariance relationships in the resulting registered functions are optimal according to this prior information.

For the registration and factor analysis model, we would like to use separate precision matrices in the prior on the registered functions to penalize registered function estimates for variation in directions other than 1) a scaling of the first factor, $\mathbf{f}_1$, and 2) a linear combination of the first

and second factors, $\mathbf{f}_1$ and $\mathbf{f}_2$. Thus, we again utilize the precision matrix, $\mathbf{\Sigma}^{-1}$, that is designed to penalize all variation from a given mean function and require the prior for the approximated registered functions, $\mathbf{X}_i(\mathbf{h}_i)$ to have the following property,

$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \propto$

$$exp \quad \big[-\frac{1}{2}\big((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1))'\gamma_1\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1))\big)\big] * \tag{4}$$

$$exp \quad \big[-\frac{1}{2}\big(\mathbf{X}_i((\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + z_{2i}\mathbf{f}_2))'\gamma_2\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + z_{2i}\mathbf{f}_2))\big)\big] \tag{5}$$

where $\gamma_1 > \gamma_2$ so that variation in the registered functions in directions other than a scaling of the first factor (4) is penalized more heavily than variation in directions other than a linear combination of both factors (5) (where both penalties account for vertical shifts). A specific definition of $\mathbf{\Sigma}$ can be found in Appendix A.1.

After rearranging terms and determining the appropriate normalizing constant, this criterion results in prior distribution (3) for the registered functions.

2.2 Model Specifications

The full data and prior distributions for the registration and factor analysis model assuming unregistered functions $X_i(t)$, have been observed over $\mathbf{t} = (t_1 \dots t_p)'$ are

$$\begin{aligned}
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \quad i = 1, \dots, N \quad (6)\\
\mathbf{\Sigma} &= \mathbf{P}_1 + \mathbf{P}_2 \quad (7)\\
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j}(t_k - t_{k-1})e^{w_i(t_{k-1})} \quad i = 1, \dots, N \quad j = 1, \dots, p\\
\mathbf{w}_i &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\mathbf{\Sigma} + \lambda_w^{-1}\mathbf{P}_w)\mathbb{1}\{t_1 + \sum_{k=2}^{p}(t_k - t_{k-1})e^{w_i(t_{k-1})} = t_p\} \quad (8)\\
&\quad i = 1, \dots, N\\
\mathbf{P}_w &= \mathbf{P}_2 \quad (9)\\
z_{0i} \mid \sigma^2_{z0} &\sim N(0, \sigma^2_{z0}) \quad i = 1, \dots, (N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i}\\
\sigma^2_{z0} &\sim IG(a, b)\\
z_{1i} \mid \sigma^2_{z1} &\sim N(1, \sigma^2_{z1}) \quad i = 1, \dots, N\\
\sigma^2_{z1} &\sim IG(a, b)\\
z_{2i} \mid \sigma^2_{z2} &\sim N(0, \sigma^2_{z2}) \quad i = 1, \dots, N\\
\sigma^2_{z2} &\sim IG(a, b)\\
\mathbf{f}_1 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \quad (10)\\
\mathbf{f}_2 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \quad (11)\\
\mathbf{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2 \quad (12)\\
\eta_f &\sim G(c, d)\\
\lambda_f &\sim G(c, d)
\end{aligned}$$

Note: For simplicity, here we include only the finite dimensional representation of all functional model specifications. Keep in mind that all multivariate normal distributions above are derived from a Gaussian process distribution evaluated over $(t_1, \dots, t_p)'$. Throughout this paper we will refer to the functions and finite representations of the functions interchangeably.

In this model, a, b, c, and d are hyper-parameters defining uninformative priors on the variance components and smoothing parameters. The parameters, $z_{0i}, i = 1, \dots N$, allow the registered

functions to vary by vertical shifts from a linear combinations of the two factors, $f_1(t)$ and $f_2(t)$. The constraint, $z_{0N} = -\sum_{i=1}^{N-1} z_{0i}$, ensures the average vertical shift is estimated to be 0. The parameters, z_{1i} and $z_{2i}, i = 1, \ldots N$, are the function specific weights for $f_1(t)$ and $f_2(t)$, respectively.

In this paper, we will refer to the functions, $w_i(t)$, $t \in \mathcal{T}$, from which the warping functions, $h_i(t)$, $t \in \mathcal{T}$, are derived, as the base functions. The base functions are non-parametrically specified for optimal registration. We, however, impose the following restrictions on the warping functions:

1. $h(t_1) = t_1$
2. $h(t_p) = t_p$
3. if $t_k > t_j$, then $h(t_k) > h(t_j)$ for all $t_k, t_j \in \mathcal{T}$

Restrictions 1 and 3 are built into the definition of $h_i(t)$. Restriction 2 is imposed through the indicator function in the expression for the prior defined for each base function, $w_i(t)$, (8). Furthermore, note that $w_i(t) = 0$ corresponds to the identity warping, $h_i(t) = t$. The penalty matrix $\mathbf{\Sigma}^{-1}$ is utilized again in the prior for the base functions to penalize variation from the identity warping, with corresponding registration parameter γ_w. This penalty is necessary to avoid losing important features in each function due to extreme differences between registered and observed time. Additionally, $\mathbf{P}_w$ is a matrix designed to penalize the second squared derivative of the base functions with corresponding parameter λ_w. This penalty is not always necessary but is included to allow for additional flexibility in penalizing significant departures from the identity in the warping functions. Here we will elaborate on not only this covariance specification, but the covariance specifications for all functional parameters.

In the above model specifications, all covariance matrices are the evaluation over a finite grid of time points of a covariance function composed of a linear combinations of two bi-variate functions, $P_1(s,t)$ and $P_2(s,t)$. $P_1(s,t)$ penalizes variation in constant and linear functions and $P_2(s,t)$ penalizes function variability in all other directions. Together they define a proper covariance function. For each covariance matrix above, the specification of the registration and smoothing parameters indicate the extent the two different types of variability should be penalized for each function. For example, for both the registered functions and the base functions, we want to penalize variation in *any* direction other than that of the mean function. The covariance specifications of

$(\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}$ and $\gamma_w^{-1}\mathbf{\Sigma}$ reflect these penalties, where the magnitude of the penalty is controlled by registration parameters, γ_1, γ_2, and γ_w, (distributional assumptions 6,7, and 8). We can use $P_2(s,t)$ to penalize roughness in a given function. Here we would like the factors, $f_1(t)$ and $f_2(t)$, and the base functions to be smooth. This is achieved by the inclusion of $\lambda_f^{-1}P_2(s,t)$ and $\lambda_w^{-1}P_2(s,t)$ in the priors for these functions (distributional assumptions 10, 11, 12, 8, and 9) where the level of the penalty is controlled by the smoothing parameters λ_f and λ_w. The inclusion of $\eta_f^{-1}P_1(s,t)$ in the covariance specification for $f_1(t)$ and $f_2(t)$ is needed to define a proper covariance function for these distributions where η_f only needs to be large enough to ensure stability in the model. Note, η_f and λ_f are considered as additional unknown parameters to be estimated through the model. For the exact definitions of $P_1(s,t)$ and $P_2(s,t)$, see Earls and Hooker [3].

Short runs of the adapted variational Bayes algorithm introduced in Earls and Hooker [4] can be used to establish optimal registration parameters in this model. General guidelines include setting $\gamma_2 < \gamma_1$, where γ_1 is at least a factor of 10 larger than γ_2.

In addition to allowing more flexibility in the shape of the registered functions, a bi-product of this analysis is the estimation of the two functional directions, $f_1(t)$ and $f_2(t)$, and the associated weights of these two factors for each function, z_{1i} and z_{2i}, $i = 1, \ldots, N$, respectively. These factors tend to have a more interpretable shape than principal components, and estimating the weights for each function provides a way to group registered functions. Examples are found in both Section 3 and Section 4 .

As is typical with hierarchical models, all parameters can be estimated using MCMC samples from the joint posterior distribution. However, obtaining these samples in high-dimensional models can be expensive and time-consuming. In Earls and Hooker [4] we define and establish convergence properties for an adapted version of variational Bayes that can also be utilized here. Appendices A and B contain all of the model specifications, full-conditionals for a MCMC sampler, and details of the adapted variational Bayes algorithm.

We note here that for many analyses it is desirable for registered time, t, to correspond to the average of the estimated warping functions over the sample. In other words for all $t \in \mathcal{T}$, $\overline{h_\cdot(t)} = t$. While this model does not inherently impose this restriction, it is straightforward to shift all estimated functions, post-registration, so that this requirement is satisfied. For details on how to perform this final adjustment see Earls and Hooker [4].

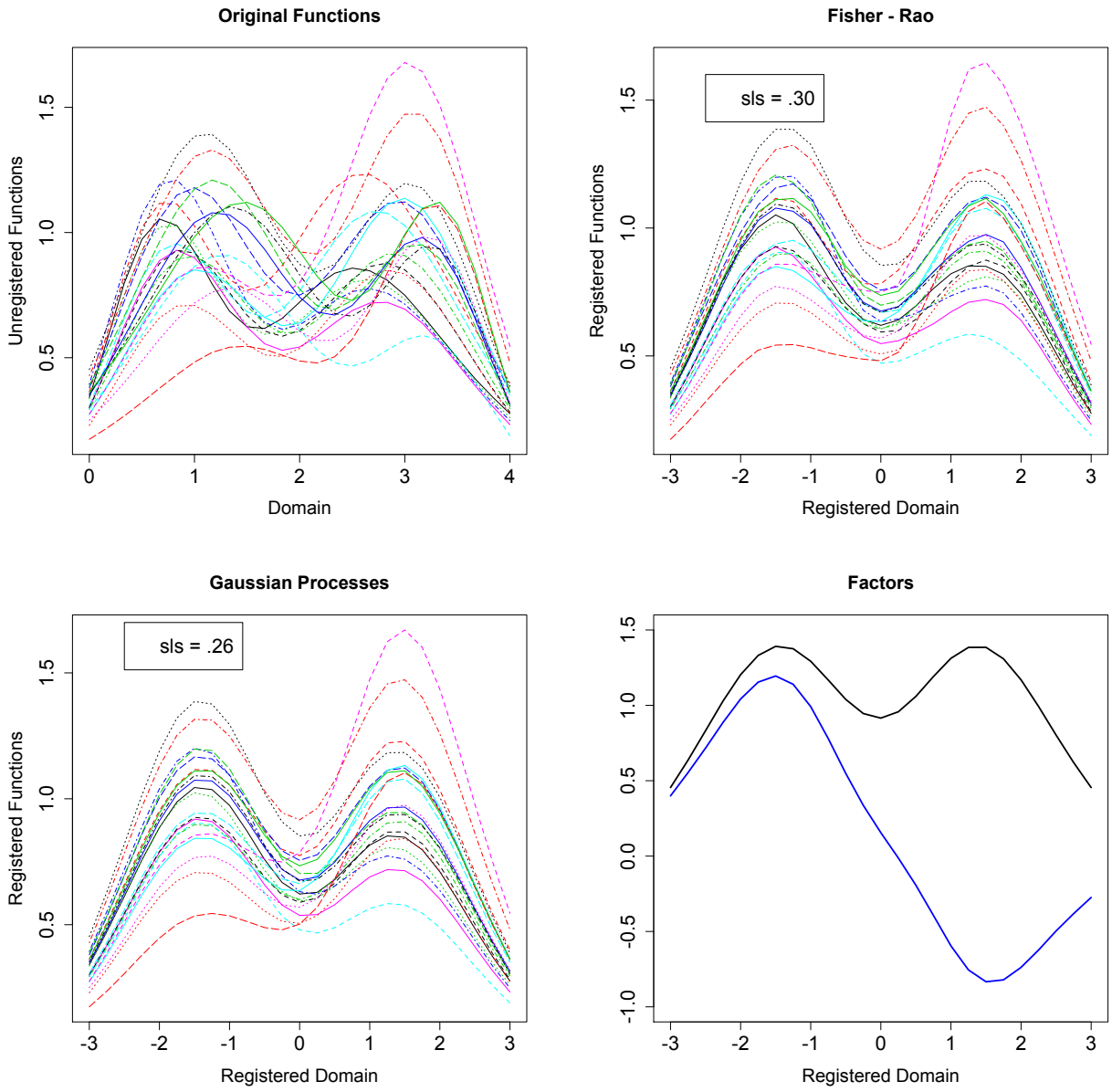

Figure 3: First Simulated Data Set. **Top Left** Original unregistered functions. **Top Right** Functions registered by F-R (R package 'fdasrvf'). **Lower Left** Functions registered by the FA model. **Lower Right** Estimated factors $\mathbf{f}_1$ and $\mathbf{f}_2$.

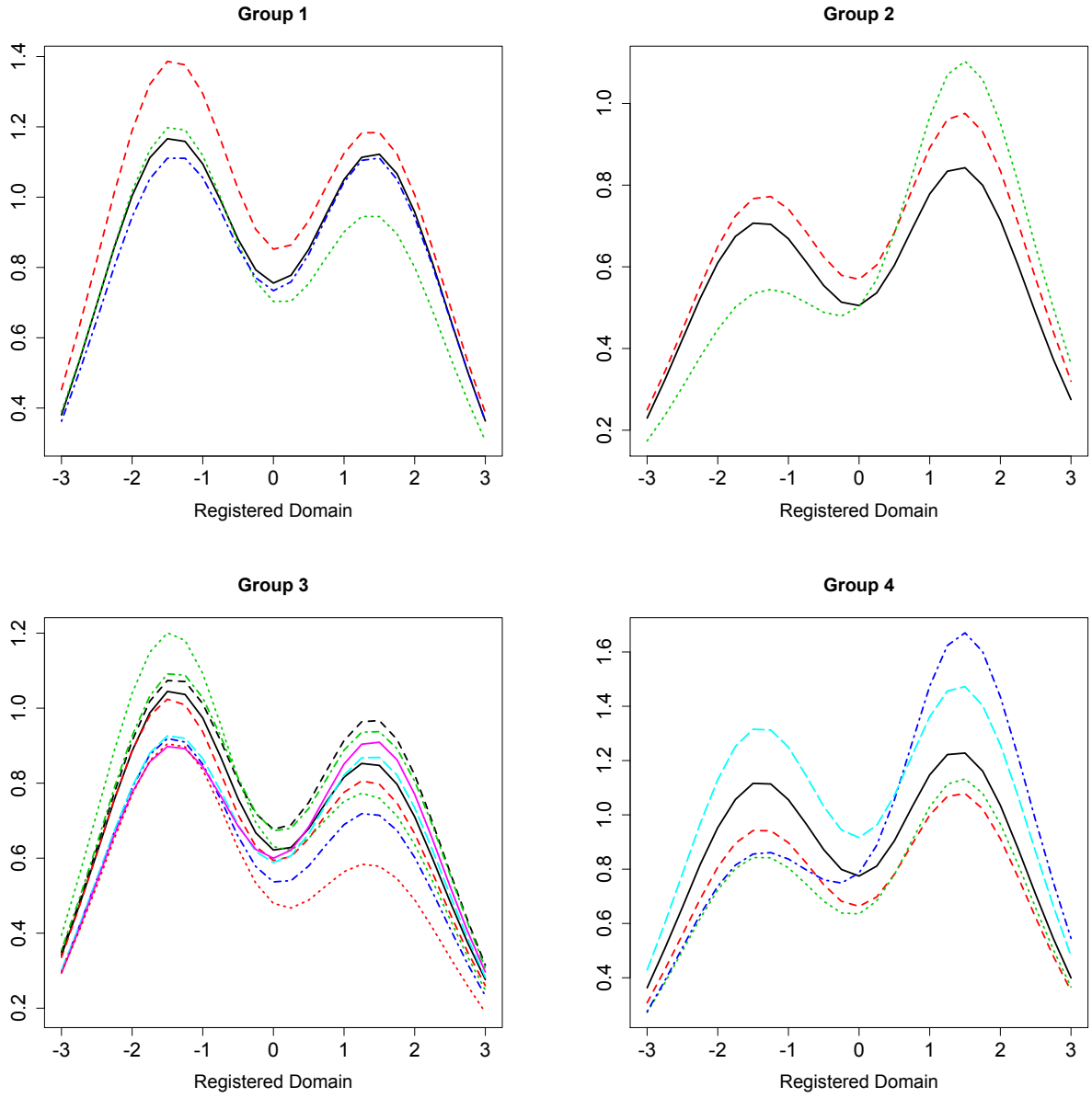

Figure 4: Four groups determined by the centered weights, $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_2$. **Top Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \tilde{z}_{2i} > 0\}$. **Top Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \tilde{z}_{2i} < 0\}$ **Lower Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \tilde{z}_{2i} > 0\}$ **Lower Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \tilde{z}_{2i} < 0\}$

3 COMPARISON TO CURRENT METHODS

One of the best registration models currently available is that proposed by Srivistava, et.al. [17]. In their work, the authors build a registration model based on the Fisher-Rao Riemannian metric that is superior to many previously considered algorithms (F-R method).

In Earls and Hooker [4], we obtain registration results similar to the F-R method using a Gaussian process model (GP). The extension of this model proposed in this paper improves on the F-R method for certain types of data. Here, we compare the registration results of F-R and of our GP model using two simulated data sets.

The Sobolev Least Squares (*sls*) criterion is used to compare the functions registered using the GP model to those registered by F-R. This criterion compares the total cross-sectional variance of the first derivatives of the registered functions to that of the original functions. Explicitly,

$$sls = \frac{\sum_{i=1}^{N} \int (X_i'(h_i(t)) - \frac{1}{N}\sum_{j=1}^{N} X_j'(h_j(t)))^2 dt}{\sum_{i=1}^{N} \int (X_i'(t) - \frac{1}{N}\sum_{j=1}^{N} X_j'(t))^2 dt} \tag{13}$$

In Srivistava, et.al. [17], *sls* is seen as the best measure of alignment in comparison to two other criterion, a least squares criterion and a pairwise correlation criterion. Lower values of *sls* correspond to better function alignment.

First Simulated Data Set The 21 unregistered functions are simulated using the algorithm originally proposed by Kneip and Ramsay [8] where the authors also consider registration in the context of multiple directions of functional variation. The registered functions $X_i(h_i(t))$, $i = 1, \ldots, 21$, are defined as $X_i(h_i(t)) = c_{1i}e^{-.5(t-1.5)^2} + c_{2i}e^{-.5(t+1.5)^2}$, $t \in [-3, 3]$ where c_{1i} and c_{2i} are iid $N(1, .25^2)$. These functions are then warped so that $h_i(t) = 6\big(\frac{e^{a_i(t+3)/6}-1}{e^{a_i}-1}\big) - 3$ if $a_i \neq 0$, where $a_i, i = 1, \ldots, 21$ are equally spaced between -1 and 1. If $a_i = 0$, $h_i(t) = t$.

Data simulated in the same way are also registered using the F-R method in Srivistava, et.al. [17]. Here we again use their method to register the simulated unregistered functions for comparison purposes. In Figure 3 are plots of the simulated unregistered functions and the functions registered using both the F-R algorithm and the proposed GP model. Both methods achieve a high degree of alignment with the GP model performing slightly better in respect to the *sls* criterion. The lower left frame of Figure 3 contains the two estimated factors to which these data are registered.

While the GP model performs similarly to F-R in this example, the added benefit of using the GP model is that the registered functions can be grouped according to their associated weights, $\mathbf{z}_1$ and $\mathbf{z}_2$ on each of the factors, $\mathbf{f}_1$ and $\mathbf{f}_2$. For functions that require registration to adequately describe the variability in the prominent features of the functions, attempting to classify functions with similar characteristics before registration will result in misclassifications that are a byproduct of the principal components or factors of the unregistered functions not reflecting the important differences that exist in these functions. In our model, variability in the weights, $\mathbf{z}_1$ and $\mathbf{z}_2$ reflect variability in significant features of the original functions without phase distortion. The result is that functions with similar weights reflect functions with similar features. In Figure 4, the registered functions are grouped by the estimated centered weights $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_2$; all functions whose centered weights lie in the same quadrant are grouped together.

Second Simulated Data Set Here we consider data with features that are not aligned well using traditional definitions of registration. Each of the 20 simulated registered functions is composed of a linear combination of two factors which is then subjected to a random warping to obtain a simulated unregistered function. The factors, $\mathbf{f}_1$ and $\mathbf{f}_2$, from which these data are simulated are found in Figure 5.

The alignment of these functions using the GP model is again compared to that obtained by F-R. For this example, the quality of alignment is best assessed by using the Sobolev Least Squares criterion separately for each of two groups of functions. Group 1 consists of functions for which $\hat{z}_{2i} > 0$. The second group is characterized by functions for which $\hat{z}_{2i} < 0$. The final *sls* value is the sum of the *sls* values for the two groups.

In Figure 5 are plots of the simulated unregistered functions, the functions registered by F-R, and the functions registered by GP. Not only is the *sls* value lower for the GP model, visually it is apparent that functions registered by the GP model are better aligned. In this example the estimated factors closely resemble the original factors from which the data are simulated. These can be seen in Figure 6. Also, in Figure 6 are three groups of registered functions determined only by classifying the estimated weights on the second factor.

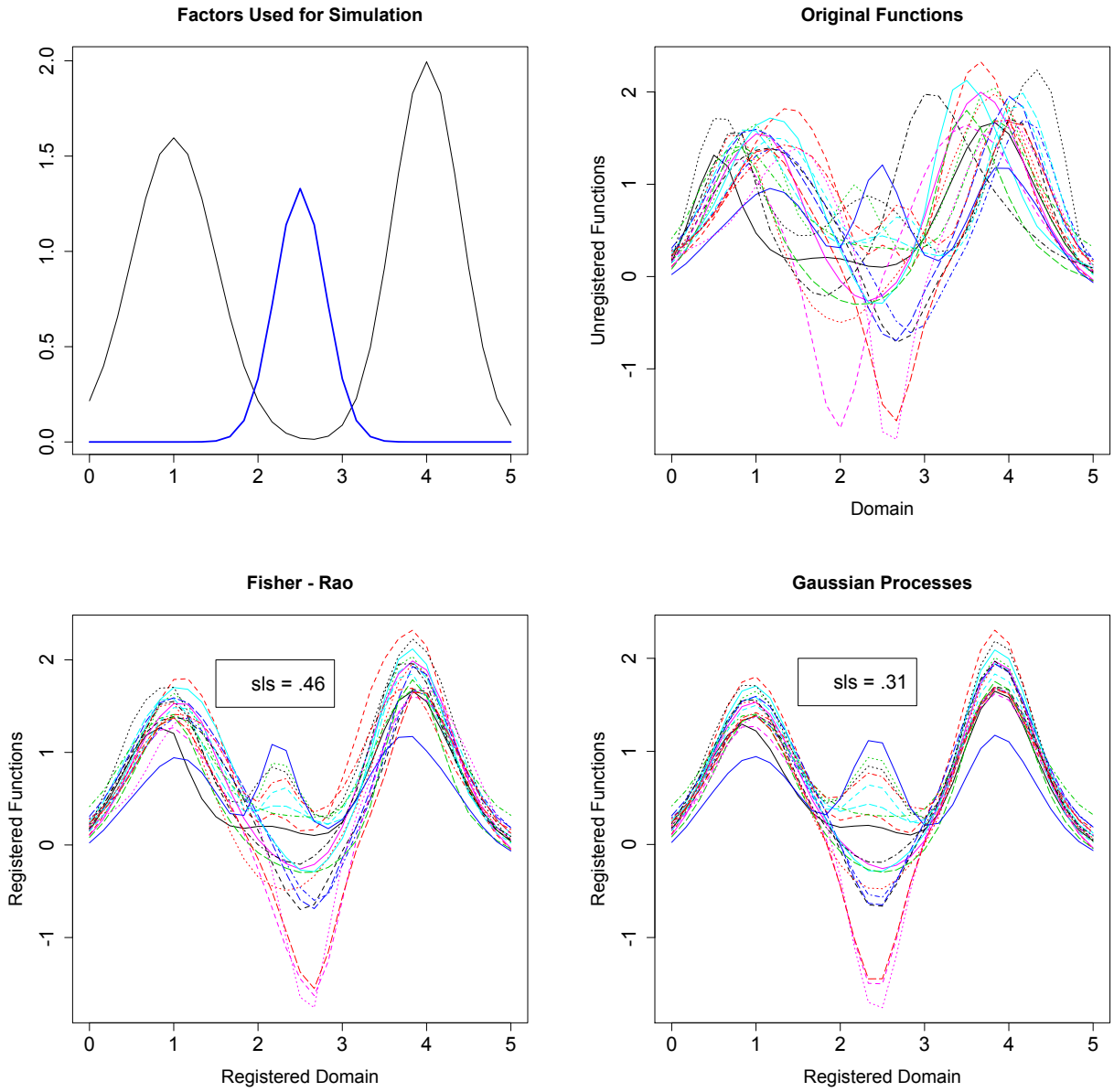

Figure 5: Second Simulated Data Set. **Top Left** The two factors used to simulate data before warping. **Top Right** Simulated unregistered functions. **Lower Left** Functions registered by F-R (R package 'fdasrvf'). **Lower Right** Functions registered by the GP model.

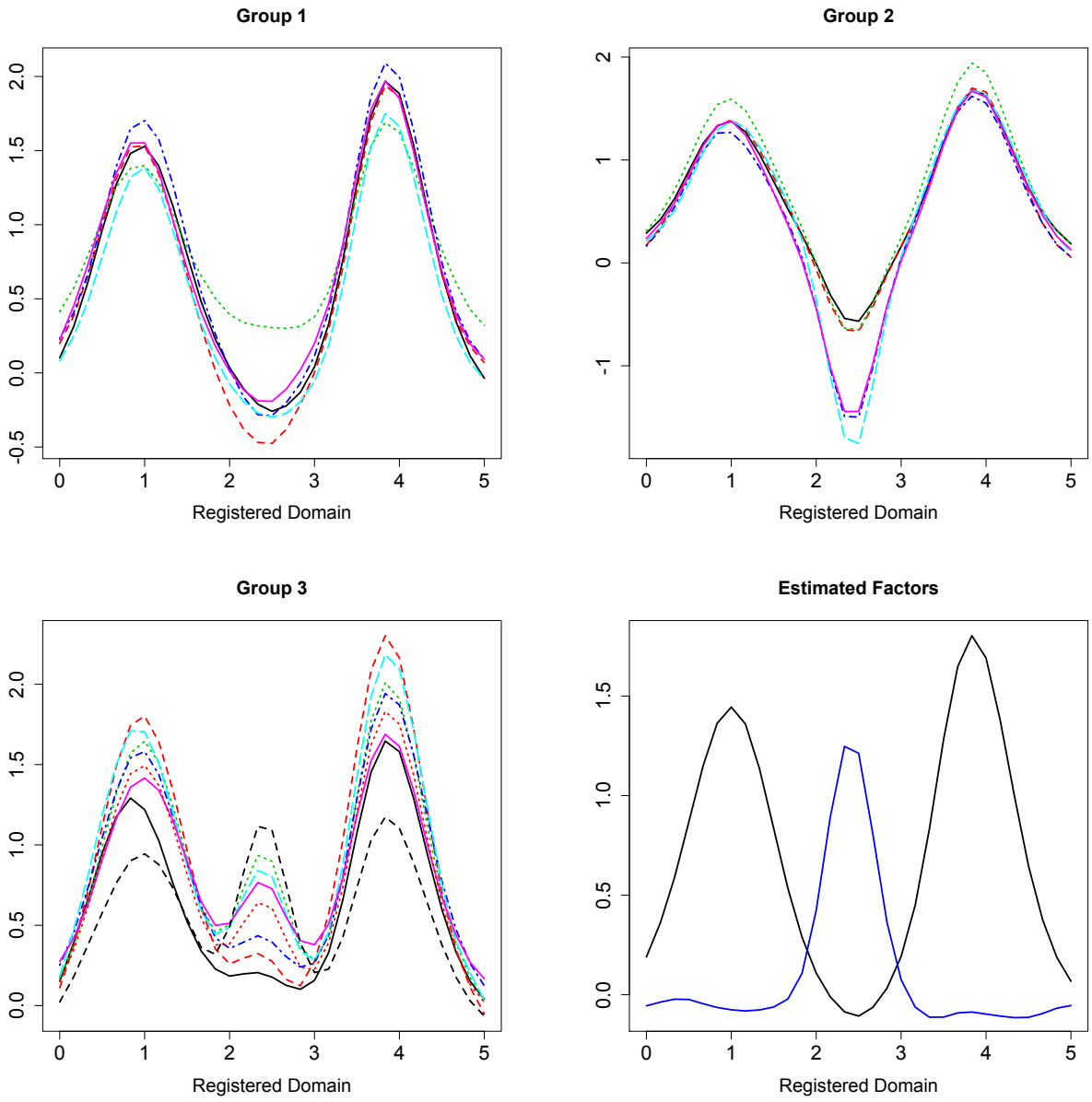

Figure 6: Three groups determined by the estimated weights on the second factor, $\mathbf{z}_2$. **Top Left** $\{X_i(h_i(t)) : \hat{z}_{2i} \in [-.1, .1]\}$. **Top Right** $\{X_i(h_i(t)) : \hat{z}_{2i} < -.1\}$ **Lower Left** $\{X_i(h_i(t)) : \hat{z}_{2i} > .1\}$ **Lower Right** Estimated factors, $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, determined by the GP model.

4 THE JUGGLING DATA: REGISTRATION AND GROUPING

The juggling data consist of three different functional data sets obtained by recording the finger position of Dr. Michael Newton (Biostatistics, University of Wisconsin) as he juggles. These data were collected in collaboration with Dr. James Ramsay (Psychology, McGill University), Dr. David Ostry (Psychology, McGill University), and Dr. Paul Gribble (Psychology, University of Western Ontario) and can be downloaded from **http://mbi.osu.edu/programs/current_topic_workshop/** under the link for the 2012 workshop, *CTW: Statistics of Time Warpings and Phase Variations*. As Dr. Newton juggled, the following were recorded: 1) the horizontal position of the right forefinger in the frontal plane, 2)the horizontal position of the right forefinger in the sagittal plane, and 3) the vertical position of the right forefinger. For this data analysis, the first functional data set of the horizontal position of the right forefinger in the frontal plane is used to demonstrate functional data registration and grouping using our Gaussian process model. Additional information on this data set can be found in Ramsay and Silverman [13].

Description of the Juggling Data For this analysis, our observations consist of individual cycles. In each cycle, the right hand cycles smoothly oscillates from left to right as the ball is caught and released. Each functional observation begins at the apex of each cycle that corresponds to the X-coordinate of the juggler's right forefinger immediately after releasing the ball with his right hand. From here, each function takes a sharp dip as the juggler's hand moves to the left to catch the next ball. Variation in the X-coordinate of these cycles correspond to the adjustments made by the juggler after the initial movement to the left to account for differences between where the ball actually descends and where the juggler anticipates it to be. Of approximately 100 cycles available, we randomly selected 25 to use for this analysis. All cycles are considered over a common time domain ranging from 0 to 675 milliseconds where the original data are recorded in 5 millisecond intervals. Thinning the data does not significantly alter its shape, and the final data contains 28 records per functional observation (cycle) taken every 25 milliseconds.

The goal of this analysis is two-fold. The first aim is to align the prominent features in these 25 cycles in conjunction with estimating the two primary factors of which these data are composed. Secondly, using the estimated weights, $\hat{\mathbf{z}}_\mathbf{1}$ and $\hat{\mathbf{z}}_\mathbf{2}$, classify these functions into groups of functions

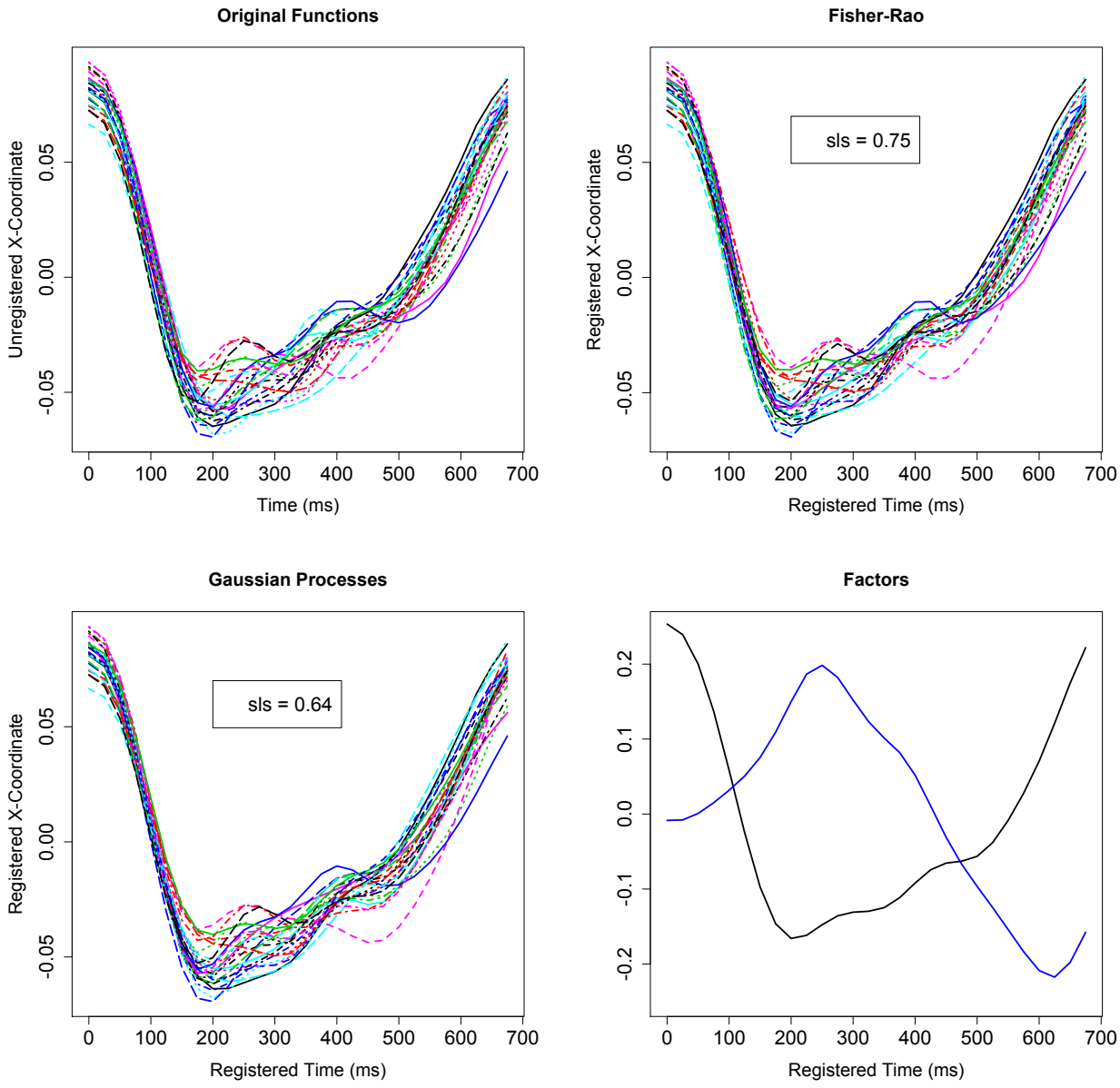

Figure 7: Juggling Data. **Top Left** Original unregistered functions. **Top Right** Functions registered by F-R (R package 'fdasrvf'). **Lower Left** Functions registered by the FA model. **Lower Right** Estimated factors, $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, determined by the GP model.

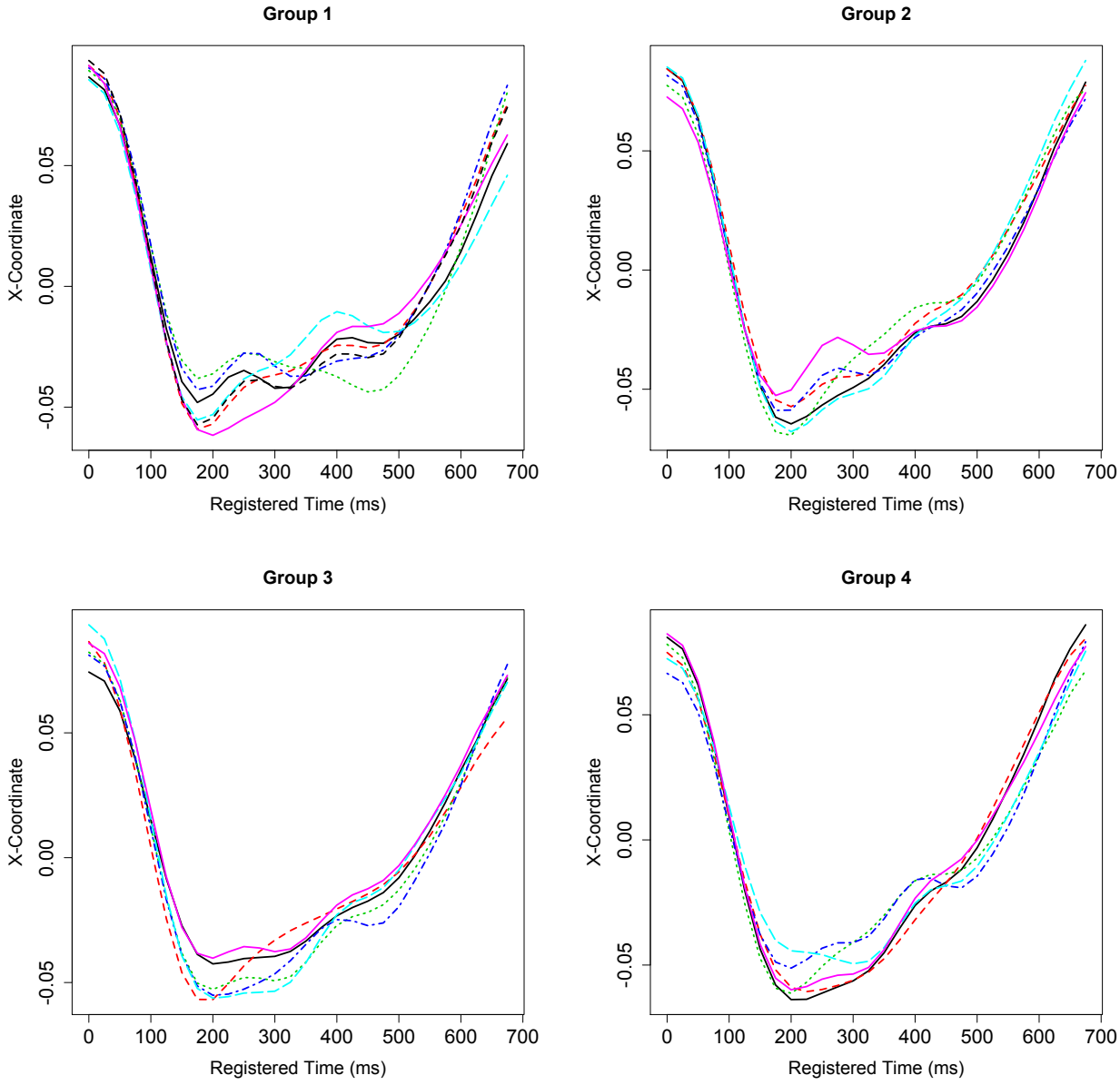

Figure 8: Four groups determined by the centered weights on the first factor, $\tilde{\mathbf{z}}_1$, and the unadjusted weights on the second factor, $\hat{\mathbf{z}}_2$. **Top Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \hat{z}_{2i} > 0\}$. **Top Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \hat{z}_{2i} < 0\}$ **Lower Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \hat{z}_{2i} > 0\}$ **Lower Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \hat{z}_{2i} < 0\}$

that share similar features.

Figure 7 contains plots of the unregistered functions, the functions registered by F-R, the functions registered by GP and the first two estimated primary directions of variation in these functions. Here again, based on the *sls* criterion, the GP model provides a better function alignment than F-R.

The estimated registered functions are split into four groups based on the estimated weights for each function in each of the primary directions of variation. Since all estimated weights on the first factor were positive, we centered these weights to delineate functions with large weights on the first factor and those with smaller weights on the first factor. In contrast, the variation in the estimated weights on the second factor could be described by whether this weight was positive or negative. Four groups were determined by the magnitude of the estimated weight on the first factor and the sign of the estimated weight on the second factor. This is equivalent to grouping by the quadrant in which the centered weight on the first factor and the unadjusted weight on the second factor, $(\tilde{z}_{1i}, \hat{z}_{2i})$, lays.

Figure 8 contains the resulting four groups of functions. In cycles with large weights on the first factor, found in the top two plots, the peak found in these functions between 200 and 300 milliseconds corresponds to the juggler overcompensating for moving his hand too far to the left to catch the ball by making a sharp movement to the right. This is then followed by another adjustment to the left. A positive weight on the second factor in the first group corresponds to cycles where the juggler needs to make another significant adjustment in his hand position between 300 and 500 milliseconds. This not only results in another significant peak in these functions, but also corresponds to the juggler releasing the ball with his hand further to the left when compared to the previous cycle. This can be seen in these functions having a smaller X-coordinate at the end of the cycle than at the beginning of the cycle. Differences in the X-coordinate when the ball is released between the previous and current cycle are much less prominent in Group 2. Groups 3 and 4 contain cycles with smaller estimated weights on the first factor. This is reflected in more subtle adjustments in hand position in the horizontal plane between 200 and 300 milliseconds where these functions stay relatively flat. Again, as seen in Groups 1 and 2, here we see that the sign of the weight on the second factor delineates between cycles where there is a distinct change in the X-coordinate at the time the ball is released between the previous and current cycle and

when there was not. Group 3 corresponds to cycles where the ball is released further to the left in comparison to the release point for the previous cycle while those in Group 4 contain cycles with similar release points to that of the previous cycle.

This example illustrates how our model for registration and factor analysis uncovers functional differences and similarities that cannot be detected as effectively using traditional methods for registering functional data. Furthermore, since the weights on each factor for each function are additional unknowns in our model, we can quantify how certain we are a function belongs to a particular group by looking at the variability in the posterior sample of these weights.

5 DISCUSSION

In this paper, we have proposed a Bayesian hierarchical model for functional registration and factor analysis. This model reduces phase variability in functions that when registered have significant variation in more than one primary functional direction. We have shown for these types of functions, our registration model outperforms one of the best registration algorithms available. Furthermore, in addition to performing functional registration, with our model two primary directions of variation are estimated for the registered functions. Each registered function is primarily composed of a weighted combination of these factors, and by classifying the estimated weights on these factors, functions can easily be grouped.

For this analysis, a Metropolis within Gibbs sampler is used to obtain MCMC samples from the joint posterior distribution of all parameters. In general, MCMC sampling is inefficient for high-dimensional models. However, in this particular model, reasonable estimates can be obtained by utilizing an adapted form of variational Bayes that significantly reduces computational costs. If MCMC sampling is preferred, more efficient sampling schemes are available for use. In particular, Calderhead [2] suggests that population MCMC can be employed to allow both global and local movement throughout the parameter space for a more efficient sampler and could be applied here. Initial estimates for an MCMC sampler should be obtained using AVB for optimal performance.

This work was possible through the flexibility of prior assumptions in a Bayesian hierarchical model. We have shown for functional data these models can encompass a multitude of inferential procedures including latent function estimation, functional linear regression, functional registration, and functional registration with factor analysis (Earls and Hooker [3] and Earls and Hooker

[4]). This list however is not exhaustive, and, in general, combinations of these inferential procedures can be encompassed within a single model. For instance, in Earls and Hooker [4] we propose a model for registering latent functions. Another example might be to encompass functional linear regression and registration within the same model. The advantage to these types of models are that common pre-processing steps such as smoothing, registration, and covariance estimation can be included within the model so that uncertainty in these steps can be encompassed in the final inferential procedure. Future work will focus on exploring further extensions to these models and continuing to pursue greater computational efficiency for these models.

Acknowledgements

This research was partially supported from NSF grants DEB-0813743, CMG-0934735 and DMS-1053252.

APPENDIX A

Below, in detail, are the specifications for the hierarchical Bayesian registration and factor analysis model discussed in this paper. The first section includes the basic model for functional data registration and factor analysis also found in Section 2. Section A.2 describes the MCMC sampling scheme for this model.

A.1 Factor Analysis

As discussed in Section 2, the initial assumption of this model is that we are interested in registering and possibly grouping functional data, $X_i(t), i = 1, \ldots, N$. The registered functions, $X_i(h_i(t))$, $i = 1 \ldots N$, are assumed to be characterized almost completely by a linear combination

of two factors, $f_1(t)$ and $f_2(t)$. Below are the data and prior distributions used for this model.

$$
\begin{aligned}
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \quad i = 1, \ldots, N \\
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j} (t_k - t_{k-1}) e^{w_i(t_{k-1})} \quad i = 1, \ldots, N \quad j = 1, \ldots, p \\
\mathbf{w}_i \mid \lambda_w &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\mathbf{\Sigma} + \lambda_w^{-1}\mathbf{P}_w) \mathbb{1}\{t_1 + \sum_{k=2}^{p} (t_k - t_{k-1}) e^{w_i(t_{k-1})} = t_p\} \quad i = 1, \ldots, N \\
z_{0i} \mid \sigma^2_{z0} &\sim N(0, \sigma^2_{z0}) \quad i = 1, \ldots, (N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i} \\
\sigma^2_{z0} &\sim IG(a, b) \\
z_{1i} \mid \sigma^2_{z1} &\sim N(1, \sigma^2_{z1}) \quad i = 1, \ldots, N \\
\sigma^2_{z1} &\sim IG(a, b) \\
z_{2i} \mid \sigma^2_{z2} &\sim N(0, \sigma^2_{z2}) \quad i = 1, \ldots, N \\
\sigma^2_{z2} &\sim IG(a, b) \\
\mathbf{f}_1 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{f}_2 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2 \\
\eta_f &\sim G(c, d) \\
\lambda_f &\sim G(c, d)
\end{aligned}
$$

$\mathbf{\Sigma}$ is a fixed matrix designed to penalize variation in any direction from the corresponding mean of the distribution in which it is utilized. It is composed of two matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$, such that $\mathbf{\Sigma} = \mathbf{P}_1 + \mathbf{P}_2$. $\mathbf{P}_1$ penalizes variation from the mean in constant and linear directions, and $\mathbf{P}_2$ penalizes variation from the mean in directions of curvature.

$\mathbf{P}_2$ is also used to penalize curvature in the base functions and factors, $f_1(t)$ and $f_2(t)$, with associated smoothing parameters λ_w and λ_f. Further details of the construction of $\mathbf{P}_1$ and $\mathbf{P}_2$ are found in Earls and Hooker [3].

A.2 MCMC Sampling

Using these assumptions, the following full conditional distributions are derived to run a MCMC sampler. Note, this list will not include a full conditional for the base functions or registered functions as their priors are not conjugate. Instead, the base and registered functions

are sampled via a Metropolis step.

$$
\begin{aligned}
\mathbf{f}_1 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_1|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_1|rest} &= (\sum_{i=1}^{N} z_{1i}^2(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1})^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_1|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}\Big[(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{1i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)\Big)\Big] \\
\mathbf{f}_2 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_2|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_2|rest} &= \Big(\sum_{i=1}^{N} z_{2i}^2\Big(\frac{\gamma_2^2}{\gamma_1+\gamma_2}\Big)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1}\Big)^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_2|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}\Big[\gamma_2\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{2i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1)\Big)\Big] \\
z_{0i} \mid rest &\sim N(\mu_{z_{0i}|rest}, \sigma^2_{z_{0i}|rest}) \\
\sigma^2_{z_{0i}|rest} &= (\sigma_{z_0}^{-2} + 2 * \mathbf{1}_p'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1} \\
\mu_{z_{0i}|rest} &= \sigma^2_{z_{0i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - \mathbf{X_N}(\mathbf{h_N}) + (z_{1N}-z_{1i})\mathbf{f}_1 + \Big(\frac{\gamma_2}{\gamma_1+\gamma_2}\Big)(z_{2N}-z_{2i})\mathbf{f}_2 - \\
&\quad \sum_{j=1}^{N-1} z_{0j}\mathbb{1}\{j\neq i\}\mathbf{1}_p\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p \\
\sigma^2_{z_0} \mid rest &\sim IG(a+(N-1)/2, b+1/2\sum_{i=1}^{N-1} z_{0i}^2) \\
z_{1i} \mid rest &\sim N(\mu_{z_{1i}|rest}, \sigma^2_{z_{1i}|rest}) \\
\sigma^2_{z_{1i}|rest} &= (\sigma_{z_1}^{-2} + \mathbf{f}_2'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{2i}|rest} &= \sigma^2_{z_{2i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_1 \\
\sigma^2_{z_1} \mid rest &\sim IG(a+N/2, b+1/2\sum_{i=1}^{N} z_{1i}^2) \\
z_{2i} \mid rest &\sim N(\mu_{z_{2i}|rest}, \sigma^2_{z_{2i}|rest}) \\
\sigma^2_{z_{2i}|rest} &= (\sigma_{z_2}^{-2} + \mathbf{f}_2'\frac{\gamma_2^2}{\gamma_1+\gamma_2}\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{2i}|rest} &= \sigma^2_{z_{2i}|rest}\gamma_2\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1)\Big)'\boldsymbol{\Sigma}^{-1}\mathbf{f}_2 \\
\sigma^2_{z_2} \mid rest &\sim IG(a+N/2, b+1/2\sum_{i=1}^{N} z_{2i}^2) \\
\eta_f \mid rest &\sim G(c+2, d+\frac{1}{2}tr\big((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_1^-\big)\big) \\
\lambda_f \mid rest &\sim G(c+(p-2), d+\frac{1}{2}tr\big((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_2^-\big)\big)
\end{aligned}
$$

APPENDIX B

B.1 Adapted Variational Bayes

The variational Bayes procedure described here is based on the variational methods proposed by Omerod and Wand [12] and Bishop [1]. Their proposed method optimizes a lower bound of the marginal likelihood which results in finding an approximate joint posterior density that has the smallest Kullback-Leibler (KL) distance, Kullback and Leibler [9], from the true joint posterior density.

In minimizing the KL distance between the approximate and true posterior distribution, parameters are updated by an optimization method that requires an approximate posterior density that not only factors but factors into components of known parametric forms. Suppose, $q(\boldsymbol{\theta})$ is the approximated posterior joint distribution. Then for some partition of $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$, $q(\boldsymbol{\theta}) = \prod_{k=1}^{d} q_k(\boldsymbol{\theta}_k)$, where each distribution q_k is of a known parametric form.

In our model, the Gaussian process priors for the base functions, $w_i(t)$, $i = 1, \ldots, N$, are not conditionally conjugate to the likelihood function. Therefore, the traditional variational Bayes optimization method does not apply directly since $q_k(\mathbf{w}_i)$, $i = 1, \ldots, N$ are not known parametric distributions. Thus, we propose an adapted variational Bayes algorithm.

After initializing all parameters, in each iteration, the adapted variational Bayes algorithm performs two steps. In the first step, the 'likelihood' as a function of the base functions is maximized. For this 'likelihood', all other parameters are fixed at their current values. The second step uses a traditional variational Bayes iterative scheme to update all other parameters. Specifically, assuming $\boldsymbol{\theta}_k = \mathbf{w}_k$, for $k = 1 \ldots N$, so that, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, the adapted variational Bayes algorithm is as follows:

1. Initialize $\boldsymbol{\theta}$

2. For each iteration, m, and each k, $k = 1, \ldots, N$, update the estimate for $\mathbf{w}_k$ so that $\mathbf{w}_k^{(m)} = \sup_{\mathbf{w}_k} q_k(\mathbf{w}_k \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1), \ldots, d)$

3. For each iteration, m, and each k, $k = (N+1), \ldots, d$, update q_k so that $q_k^{(m)} \propto exp[E_{(\boldsymbol{\theta}_{-k})}(log\ f(\boldsymbol{\theta}_k \mid rest)]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = 1, \ldots, d$, $j \neq k$

4. Repeat steps (2) and (3) until the desired convergence criterion is met

This algorithm is guaranteed to converge. However, convergence is not guaranteed to a global maximum, and in practice it is sometimes necessary to adjust the registration and warping penalties as the functions become registered. An unregistered function that requires a substantial amount of warping can cause convergence to a local maximum due to the small penalty on warping. The flexibility in warping allowed by this small penalty can cause the function to deform rather than register. This can be remedied in two ways. The first option might be to perform a simple initial warping for this function that prevents the optimization from falling into a local mode. The second option is to adjust the registration and warping parameters over time. Initially a stronger warping penalty is employed to prevent function deformation. Then, as the functions register, the warping penalty can be reduced to allow for a more complete registration. When initializing an MCMC sampler, the final penalties on warping and registration from the adapted variational Bayes algorithm should be used. For further information on the convergence properties of the adapted variational Bayes algorithm and an analysis of how well adapted variational Bayes estimates correspond to MCMC estimates, see Earls and Hooker [4].

Below are the approximate posterior distributions, $q_k(\boldsymbol{\theta}_k)$, $k = (N+1), \ldots, d$, for the adapted variational Bayes estimation procedure for the registration and factor analysis model. Note, the subscripts on the q distributions has been omitted. For a more thorough discussion and illustration of how the optimal q distributions are derived see Goldsmith et. al. [6].

$$
\begin{aligned}
q(\mathbf{f}_1) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_1)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}) \\
q(\mathbf{f}_2) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_2)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}) \\
q(z_{0i}) &\sim N(\mu_{q(z_{0i})}, \sigma^2_{q(z_{0i})}) \\
q(\sigma^2_{z_0}) &\sim IG(a_{q(\sigma^2_{z_0})}, b_{q(\sigma^2_{z_0})}) \\
q(z_{1i}) &\sim N(\mu_{q(z_{1i})}, \sigma^2_{q(z_{1i})}) \\
q(\sigma^2_{z_1}) &\sim IG(a_{q(\sigma^2_{z_1})}, b_{q(\sigma^2_{z_1})}) \\
q(z_{2i}) &\sim N(\mu_{q(z_{2i})}, \sigma^2_{q(z_{2i})}) \\
q(\sigma^2_{z_2}) &\sim IG(a_{q(\sigma^2_{z_2})}, b_{q(\sigma^2_{z_2})}) \\
q(\eta_f) &\sim G(c_{q(\eta_f)}, d_{q(\eta_f)}) \\
q(\lambda_f) &\sim G(c_{q(\lambda_f)}, d_{q(\lambda_f)})
\end{aligned}
$$

The approximate joint posterior distribution of all parameters except the base functions is

$$
q(\boldsymbol{\theta}) = \prod_{k=(N+1)}^{d} q_k(\boldsymbol{\theta}_k) = q(\mathbf{f}_1)q(\mathbf{f}_2)q(\sigma^2_{z_0})q(\sigma^2_{z_1})q(\sigma^2_{z_2})q(\eta_f)q(\lambda_f) \prod_{i=1}^{(N-1)} q(z_{0i}) \prod_{i=1}^{N} q(z_{1i})q(z_{2i}) \qquad (14)
$$

As the q densities are all of known distributional forms, updating these densities is equivalent to updating their parameters. For each iteration, the following parameters are updated for the q densities found in (14). These updates are listed in an order that allows the convergence criterion to be calculated. Further details on the convergence criterion can be found in Appendix B.2.

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} = \Big[\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^-\Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_1)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N}\mu_{q(z_{1i})}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)})\big)\Big]$$

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} = \Big[\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\frac{\gamma_2^2}{\gamma_1+\gamma_2}\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^-\Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_2)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N}\mu_{q(z_{2i})}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)})\big)\Big]$$

$$\sigma^2_{q(z_{0i})} = (\mu_{q(\sigma^{-2}_{z_0})} + \mathbf{1}_p'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1}$$

$$\mu_{q(z_{0i})} = \sigma^2_{q(z_{0i})}\big(\mathbf{X}_i(\mathbf{h}_i) - \mathbf{X}_N(\mathbf{h}_N) + (\mu_{q(z_{1N})} - \mu_{q(z_{1i})})\mu_{q(\mathbf{f}_1)} + \frac{\gamma_2}{\gamma_1+\gamma_2}(\mu_{q(z_{2N})} - \mu_{q(z_{2i})})\mu_{q(\mathbf{f}_2)}\big) -$$
$$\sigma^2_{q(z_{0i})}\big(\sum_{j=1}^{N-1}\mu_{q(z_{0j})}\mathbb{1}\{i\neq j\}\mathbf{1}_p\big)$$

$$\sigma^2_{q(z_{1i})} = (\mu_{q(\sigma^{-2}_{z_1})} + tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{1i})} = \sigma^2_{q(z_{1i})}\Big(\mu'_{q(\mathbf{f}_1)}(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)})\big)\Big)$$

$$\sigma^2_{q(z_{2i})} = (\mu_{q(\sigma^{-2}_{z_2})} + \frac{\gamma_2^2}{\gamma_1+\gamma_2}tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)})\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{2i})} = \sigma^2_{q(z_{2i})}\Big(\mu'_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}\big(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)})\big)\Big)$$

$$d_{q(\eta_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_1^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$d_{q(\lambda_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_2^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$b_{q(\sigma^2_{z_0})} = b + 1/2\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})$$

$$b_{q(\sigma^2_{z_1})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})$$

$$b_{q(\sigma^2_{z_2})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})$$

B.2 Convergence Criterion

The adapted variational Bayes algorithm is run until changes in $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w},\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big]$ are below a certain threshhold. This value can be computed in each iteration as follows.

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w},\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ (f(\mathbf{X},\mathbf{w}\mid\boldsymbol{\theta}_{-\mathbf{w}})f(\boldsymbol{\theta}_{-\mathbf{w}})) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w}\mid\boldsymbol{\theta}_{-\mathbf{w}}) + log\ f(\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w}\mid\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_1) - log\ q(\mathbf{f}_1)\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)\big] \\
&\quad + \sum_{i=1}^{(N-1)} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{0i}) - log\ q(z_{0i})\big] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{1i}) - log\ q(z_{1i})\big] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(z_{2i}) - log\ q(z_{2i})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\eta_f) - log\ q(\eta_f)\big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\lambda_f) - log\ q(\lambda_f)\big]
\end{aligned}
$$

Now looking at each piece individually,

$$
\begin{aligned}
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X},\mathbf{w}\mid\boldsymbol{\theta}_{-\mathbf{w}})\big] \\
= \;& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N}\big(log[(2\pi)^{-p/2}\mid(\gamma_1+\gamma_2)^{-1}\boldsymbol{\Sigma}\mid^{-1/2}]\big)\Big] \\
& +\ E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N}-\frac{1}{2}[(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i)-2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)+ \\
& (z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p+z_{1i}\mathbf{f}_1+\frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)]\Big] \\
= \;& \sum_{i=1}^{N}\big(log[(2\pi)^{-p/2}\mid(\gamma_1+\gamma_2)^{-1}\boldsymbol{\Sigma}\mid^{-1/2}]\big) \\
& +\Big[\sum_{i=1}^{N}-\frac{1}{2}\Big(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i)- \\
& 2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{0i})}\mathbf{1}_p-2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)}- \\
& 2\mathbf{X}_i(\mathbf{h}_i)'\gamma_2\boldsymbol{\Sigma}^{-1}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)}+ \\
& (\sigma^2_{q(z_{1i})}+\mu^2_{q(z_{1i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}+\boldsymbol{\mu}_{q(\mathbf{f}_1)}\boldsymbol{\mu}'_{q(\mathbf{f}_1)})(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1})+ \\
& (\sigma^2_{q(z_{2i})}+\mu^2_{q(z_{2i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}+\boldsymbol{\mu}_{q(\mathbf{f}_2)}\boldsymbol{\mu}'_{q(\mathbf{f}_2)})\frac{\gamma_2^2}{(\gamma_1+\gamma_2)}\boldsymbol{\Sigma}^{-1})+ \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}'_p(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_1)}+2\mu_{q(z_{1i})}\mu_{q(z_{2i})}\boldsymbol{\mu}'_{q(\mathbf{f}_1)}\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}+ \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{2i})}\mathbf{1}'_p\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big)\Big]- \\
& \Big[\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})}+\mu^2_{q(z_{0i})})+\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j\neq i\}\Big]\mathbf{1}'_p(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p
\end{aligned}
$$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_1)-log\ q(\mathbf{f}_1)\big] \;=\; & E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{p}{2}log\ 2\pi+\frac{1}{2}log\ \mid\eta_f\mathbf{P}_1^-+\lambda_f\mathbf{P}_2^-\mid\Big]- \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_1\mathbf{f}_1'(\eta_f\mathbf{P}_1^-+\lambda_f\mathbf{P}_2^-)]\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2}log\ 2\pi+\frac{1}{2}log\ \mid\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}\mid\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}tr(\mathbf{f}_1\mathbf{f}_1'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)})-\mathbf{f}_1'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big]+ \\
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}\boldsymbol{\mu}'_{q(\mathbf{f}_1)}\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big] \\
= \;& C+\frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[2log\ \eta_f\big]+\frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(p-2)log\ \lambda_f\big]- \\
& \frac{1}{2}tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)}+\boldsymbol{\mu}_{q(\mathbf{f}_1)}\boldsymbol{\mu}'_{q(\mathbf{f}_1)})(\mu_{q(\eta_f)}\mathbf{P}_1^-+\mu_{q(\lambda_f)}\mathbf{P}_2^-)\Big)- \\
& \frac{1}{2}log\ \mid\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f}_1)}\mid+\frac{p}{2}
\end{aligned}
$$

where C is a constant that does not change from one iteration to the next. Similarly,

$$\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^- \mid \Big] - \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_2\mathbf{f}_2'(\eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^-)]\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_2)} \mid \Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} tr(\mathbf{f}_2\mathbf{f}_2'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}) - \mathbf{f}_2'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}\boldsymbol{\mu}_{q(\mathbf{f}_2)}'\boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] \\
&= C + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[2log\ \eta_f\big] + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(p-2)log\ \lambda_f\big] - \\
&\quad \frac{1}{2} tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \boldsymbol{\mu}_{q(\mathbf{f}_2)}\boldsymbol{\mu}_{q(\mathbf{f}_1)}')(\mu_{q(\eta_f)}\mathbf{P}_1^- + \mu_{q(\lambda_f)}\mathbf{P}_2^-)\Big) - \\
&\quad \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \mid + \frac{p}{2}
\end{aligned}$$

where C is a constant that does not change from one iteration to the next. For $\mathbf{z}_0 = (z_{01}, \ldots, z_{0(N-1)})'$

$$\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N-1}{2} log\ 2\pi - \frac{N-1}{2} log\ \sigma_{z_0}^2 - \sum_{i=1}^{N-1} -\frac{1}{2\sigma_{z_0}^2} z_{0i}^2 + \\
&\quad \frac{N-1}{2} log\ 2\pi + \frac{N-1}{2} log\ \sigma_{q(z_{0i})}^2 + \\
&\quad \sum_{i=1}^{N-1} \frac{1}{2\sigma_{q(z_{0i})}^2}(z_{0i} - \mu_{q(z_{0i})})^2\Big] \qquad (15) \\
&= \frac{N-1}{2} log\ \sigma_{q(z_{0i})}^2 - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N-1}{2} log\ \sigma_{z_0}^2\Big] - \qquad (16) \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma_{z_0}^2})}\Big(\sum_{i=1}^{N-1}(\sigma_{q(z_{0i})}^2 + \mu_{q(z_{0i})}^2)\Big) + \frac{N-1}{2}
\end{aligned}$$

For $\mathbf{z}_1 = (z_{11}, \ldots, z_{1N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_1) - log\ q(\mathbf{z}_1)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_1} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_1}} z^2_{1i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{1i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{1i})}} (z_{1i} - \mu_{q(z_{1i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{1i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_1}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_1}})}\Big(\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\mathbf{z}_2 = (z_{21}, \ldots, z_{2N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_2) - log\ q(\mathbf{z}_2)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_2} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_2}} z^2_{2i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{2i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{2i})}} (z_{2i} - \mu_{q(z_{2i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{2i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_2}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_2}})}\Big(\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\sigma^2_{z_0}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1) log\ \sigma^2_{z_0} - b\frac{1}{\sigma^2_{z_0}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + (a_{q(\sigma^2_{z_0})} + 1) log\ \sigma^2_{z_0} + \\
&\quad b_{q(\sigma^2_{z_0})}\frac{1}{\sigma^2_{z_0}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1) log\ \sigma^2_{z_0}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_0}})} - log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_0})} + 1) log\ \sigma^2_{z_0}\big] + b_{q(\sigma^2_{z_0})}\mu_{q(\frac{1}{\sigma^2_{z_0}})}
\end{aligned}
$$

For $\sigma^2_{z_1}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_1} - b\frac{1}{\sigma^2_{z_1}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + (a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1} + \\
&\quad b_{q(\sigma^2_{z_1})}\frac{1}{\sigma^2_{z_1}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1)log\ \sigma^2_{z_1}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_1}})} - log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1}\big] + b_{q(\sigma^2_{z_1})}\mu_{q(\frac{1}{\sigma^2_{z_1}})}
\end{aligned}
$$

For $\sigma^2_{z_2}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_2} - b\frac{1}{\sigma^2_{z_2}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + (a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2} + \\
&\quad b_{q(\sigma^2_{z_2})}\frac{1}{\sigma^2_{z_2}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[-(a+1)log\ \sigma^2_{z_2}\big] - b\mu_{q(\frac{1}{\sigma^2_{z_2}})} - log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[(a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2}\big] + b_{q(\sigma^2_{z_2})}\mu_{q(\frac{1}{\sigma^2_{z_2}})}
\end{aligned}
$$

For η_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\eta_f) - log\ q(\eta_f)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \eta_f - d\eta_f - \\
&\quad log\ \frac{d^{c_{q(\eta_f)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - c\ log\ \eta_f + d_{q(\eta_f)}\eta_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d^{c_{q(\eta_X)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - 2E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ \eta_f\big] - d\mu_{q(\eta_f)} + \\
&\quad d_{q(\eta_f)}\mu_{q(\eta_f)}
\end{aligned}
$$

For λ_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\lambda_f) - log\ q(\lambda_f)\big] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \lambda_f - d\lambda_f - \\
&\quad log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - \Big(\frac{p-2}{2} + c - 1\Big)\ log\ \lambda_f + d_{q(\lambda_f)}\lambda_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - (p-2)E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ \lambda_f\big] - d\mu_{q(\lambda_f)} + \\
&\quad d_{q(\lambda_f)}\mu_{q(\lambda_f)}
\end{aligned}
$$

The expression for $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})\big]$ can be simplified much further by combining terms that cancel out. However, in some cases the ability to cancel terms depends on the order of the updates. For instance, in the expression, $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\sigma_{z_0}^2) - log\ q(\sigma_{z_0}^2)\big]$, the terms $-b\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ and $b_{q(\sigma_{z_0}^2)}\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ cancel with $-\frac{1}{2}\mu_{q(\frac{1}{\sigma_{z_0}^2})}\Big(\sum_{i=1}^{N-1}(\sigma_{q(z_{0i})}^2 + \mu_{q(z_{0i})}^2)\Big)$ from $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\big[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)\big]$ as long as the parameters of $q(\mathbf{z}_0)$ are updated before $b_{q(\sigma_{z_0}^2)}$. For convenience, we have taken account the ordering necessary to compute the convergence criterion in the updates given above. Additionally, note all components in this expression that do not change from one iteration to the next can be ignored.

References

[1] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

[2] Calderhead, B., Girolami, M., and Lawrence, N.(2009). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Advances in neural information processing systems* **21**, 217-224.

[3] Earls, C., and Hooker, G. (2014). Bayesian covariance estimation and inference in latent Gaussian process models. *Statistical Methodology* **18**, 79-100.

[4] Earls, C., and Hooker, G. (2014). Adapted Variational Bayes for Functional Data Registration, Smoothing, and Prediction. *in review.*

[5] Gervini, D., and Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society, Ser. B* **66**, 959-971.

[6] Goldsmith, J., Wand, M.P., and Crainiceanu, C.(2011). Functional regression via variational Bayes. *Electronic Journal of Statistics* **5**, 572.

[7] James, G.M.(2007). Curve alignment by moments. *The Annals of Applied Statistics* **1**, 2, 480-501.

[8] Kneip, A., and Ramsay J.O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103**, 483, 1155-1165.

[9] Kullback, S., and Leibler, D.(1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79-86.

[10] Liu, X., andMüller, H.G.(2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* **99**, 687-699.

[11] Liu, X., and Yang, M.C.K. (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis* **53**, 1361-1376.

[12] Omerod, J., and Wand, M. (2010). Explaining variational approximations. *The American Statistician* **64**, 140-153.

[13] Ramsay, J., and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

[16] Ramsay, J., and Silverman, B. (2005). *Functional Data Analysis.* Springer, New York.

[15] Ramsay, J.O., and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society.Series B (Statistical Methodology)* **60**, 2, 351-363.

[16] Ramsay, J.O., and Silverman, B. (2005). *Applied Functional Data Analysis.* Springer-Verlag, New York.

[17] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J.S. (2011). Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*

[18] Tang, R., and Müller, H.G.(2008). Pairwise curve synchronization for functional data. *Biometrika* **95**, 4, 875-889.